Shivam Wahi

Senior Data Engineer

📞 (+91) 7838284268 — ✉ shivamwahi66@gmail.com — 💼 shivamwahi66

Profile Summary

Senior Data Professional with ~**4+ years** of experience in big data ecosystems, cloud platforms (**Azure, AWS**), and enterprise data solutions. Skilled in designing scalable **ETL pipelines**, **Data Warehouse Modeling**, and migrating legacy data warehouses (Oracle/Teradata) to **Cloud**. Proficient in **Python, SQL, Apache Spark, and DBT**, with a strong foundation in big data technologies and advanced data modeling. Adept at delivering **cloud-native, insight-driven architectures** that enhance decision-making and operational efficiency.

Experience

Senior Data Engineer *June 2025 - Present*
Publicis Sapient
Project: National Australia Bank (NAB) - Banking Data Platform (Onsite Assignment)

- **Developing scalable data transformations using DBT** for NAB's banking platform, implementing modular and reusable transformation logic to ensure data quality and consistency across critical financial datasets.
- **Orchestrating complex data workflows using Apache Airflow**, designing robust DAGs for automated pipeline execution, monitoring, and error handling in production banking environments.
- **Working onsite in Australia** to collaborate directly with NAB stakeholders, ensuring seamless integration of data engineering solutions with business requirements and regulatory compliance standards.
- **Implementing data governance frameworks** for sensitive banking data, adhering to Australian financial regulations (APRA) and maintaining comprehensive data lineage documentation for audit purposes.
- **Optimizing pipeline performance** through advanced DBT techniques, including incremental models, macros, and testing frameworks to enhance data processing efficiency for high-volume banking transactions.
- **Leading cross-functional collaboration** with local business analysts, data scientists, and infrastructure teams to deliver enterprise-grade data solutions for Australia's fourth-largest bank.

Data Modeller *Sept 2024 - June 2025*
Tredence Analytics Solutions Private Limited
Project: Data Engineering for CPG Client - PepsiCo

- **Designed the Silver layer data model from scratch** for a new application ingesting data via **Azure Event Hub**, enabling real-time processing and faster business insights.

- **Modeled 16 tables for Oracle Fusion Application**, refining them to 11 based on grain analysis, ensuring scalability, standardization, and seamless integration with existing data lake tables.

- **Mapped SAP S/4 supply chain tables** to existing data lake structures, improving **demand forecast accuracy**, optimizing inventory management, and enhancing supply chain efficiency.

- **Led the migration of 20+ tables from Teradata to Azure**, leveraging **Hackolade** for schema design, reducing costs, and improving the scalability of analytics solutions.

- **Ensured data compliance** by collaborating with the **Data Governance (DG) team**, securing approvals based on security classifications, and reducing business risk.

- **Collaborated with the Data Engineering team** to define and refine transformation logic, ensuring consistency, accuracy, and optimal performance across data pipelines.

- **Streamlined CI/CD deployment** by maintaining **STTM documents** and automating DDL releases through Git, accelerating deployment cycles and improving the stability of the data pipeline.

- Leveraged SQL Server Integration Services (SSIS) for developing and automating ETL workflows, enhancing data integration and streamlining onboarding of new data sources.

Data Engineer   *Sep 2021 – Sep 2024*
Cognizant Technology Solutions India Pvt. Ltd.
Project: Molina Healthcare – Big Data Development

- Migrated critical on-premise pipelines to **Azure Databricks**, ensuring minimal downtime and improved cloud-native performance. Refactored existing Scala scripts into optimized Databricks notebooks, and automated workflows using **AutoSys** and CI/CD for multi-environment deployments.

- Developed a **generic ingestion framework in Scala** supporting incremental, full, and deletion loads with built-in auditing, replacing fragmented scripts and standardizing data ingestion across multiple source systems.

- Enhanced domain-specific pipelines including **Provider Data**, **Appeals and Grievances**, and **Clinical Records**, improving processing accuracy and business alignment.

- Built scalable ingestion pipelines using **Databricks Autoloader** for fixed-width files and developed robust workflows for semi-structured **JSON data** (e.g., Jira and ServiceNow sources).

- Contributed to an acquisition data integration initiative, seamlessly ingesting historical data from acquired entities into existing datasets to maintain continuity.

- Utilized **Cloudera-based Apache Spark** for on-prem data processing and **Hive (Parquet)** external tables for data curation and reporting in the clinical and enrollment domains.

- Automated key operations using **AutoSys** job scheduler, enhancing reliability and reducing manual intervention by 80

- Investigated and resolved business-critical data issues, conducting detailed **root cause analysis** to ensure data integrity and trustworthiness.

- Authored detailed **data dictionaries** and mapped data lineage from **SQL Server** to **HDFS/ADLS**, enabling clear business understanding of pipeline transformations.

- Provided ongoing **L2/L3 support** for Azure-migrated pipelines, ensuring business continuity and high availability of data.

- Played a leading role in building, testing, and deploying new datasets while enhancing existing ones for **enrollment, clinical, claims, and provider** data domains.

Education

Bachelor of Technology in Mechanical and Automation Engineering *Jul '18 - Jun '21*
Guru Gobind Singh Indraprastha University, Delhi
GPA: 9.49

Diploma in Mechanical Engineering *Aug '15 - Jul '18*
Govind Ballabh Pant Institute of Technology, Delhi
Percentage: 80.5%

Secondary Education *2015*
Dayanand Model School, Delhi
CGPA: 8.2

Skills

- **Programming Languages**: Python, Scala, Java, SQL

- **Database Systems**: MS SQL Server, MySQL, MongoDB, PostgreSQL

- **Cloud Platforms**: Azure (Data Factory, Databricks, Synapse), AWS (RedShift, Glue, S3), GCP (BigQuery)

- **Big Data & ETL Technologies**: Apache Kafka, Apache Spark, Hive, Snowflake, **DBT (Data Build Tool)**, **SSIS (SQL Server Integration Services)**

- **Workflow Orchestration & CI/CD**: **Apache Airflow**, Azure DevOps (Pipelines, Boards), AutoSys Job Scheduler, Git, Jenkins

- **Data Modeling & Architecture**: Hackolade, Dimensional Modeling, Data Vault, Star Schema, Snowflake Schema

- **Data Visualization & Analytics**: Power BI, Microsoft Fabric, Tableau

- **Generative AI & Machine Learning**: OpenAI API, LangChain, LLM-based automation, MLOps

- **BFSI Domain Knowledge**: Regulatory Compliance, Risk Management, Financial Data Processing

- **Methodologies**: Software Development Life Cycle (SDLC), Data Life Cycle Management, Agile/Scrum

- **Collaboration & Version Control**: Jira, Confluence, Git, GitHub, Bitbucket

Certifications & Accreditations

- Databricks Certified Generative AI Engineer Associate

- Snowflake SnowPro Core

- Databricks Certified Data Engineer Professional

- Microsoft Certified: Fabric Analytics Engineer Associate (DP-600)

- Databricks Certified Data Engineer Associate

- Microsoft Certified Azure Data Engineer Associate (DP-203)

- Microsoft Certified Power BI Data Analyst Associate (PL-300)

Personal Details

- **Date of Birth:** 20 July 2000